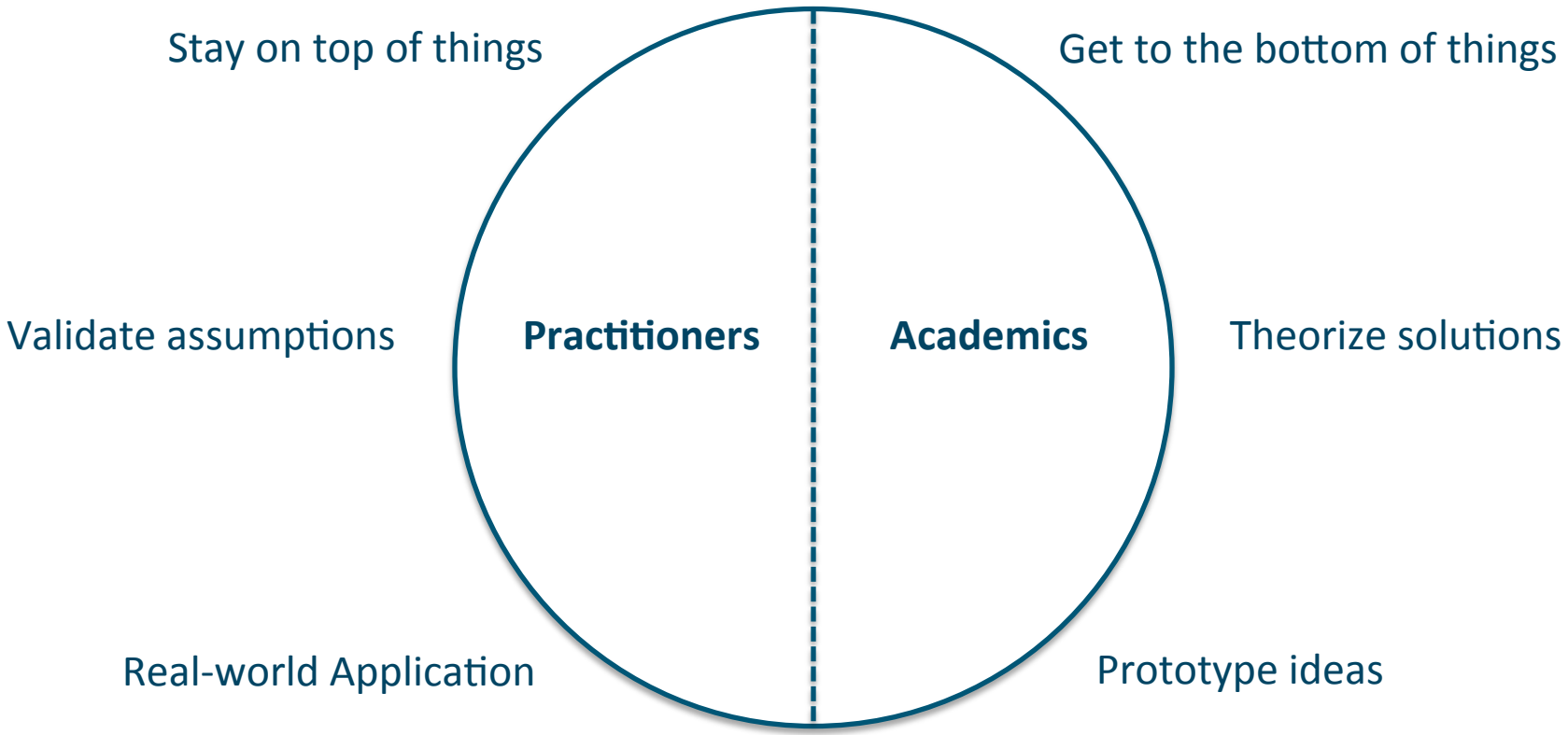
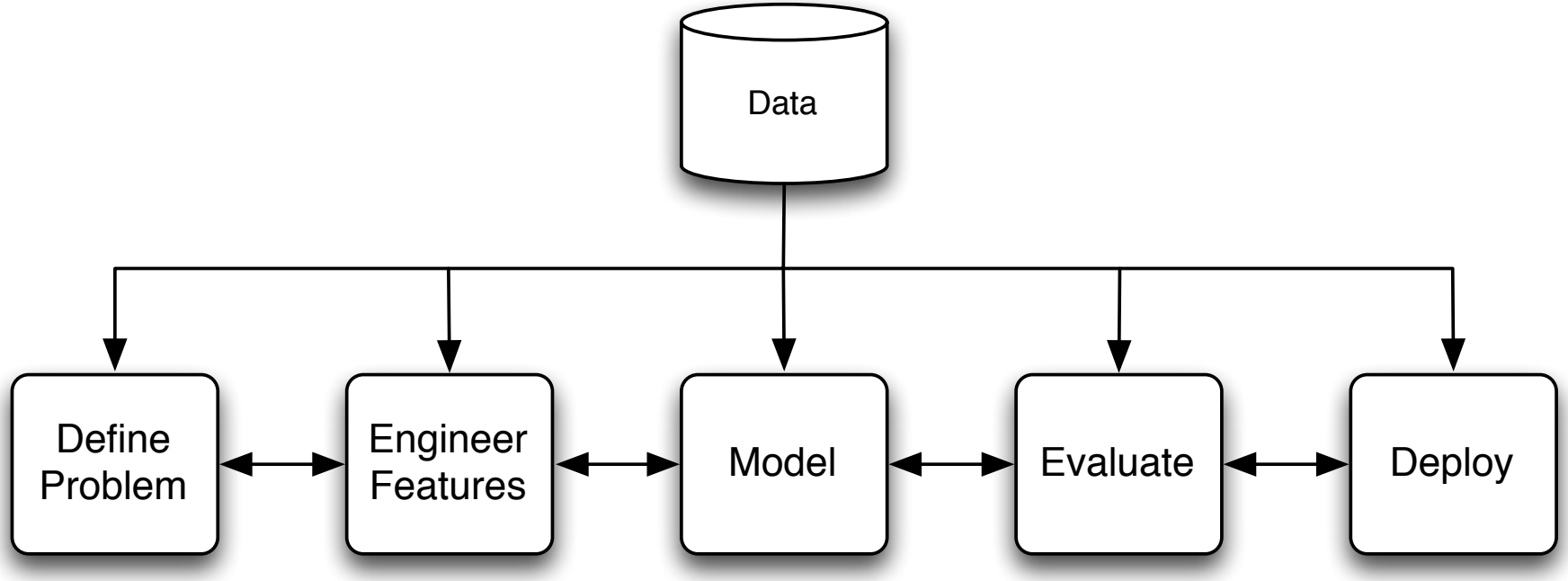
Pratical Machine Learning for Network Security

Terry Nelms
Director of Research
Damballa Labs

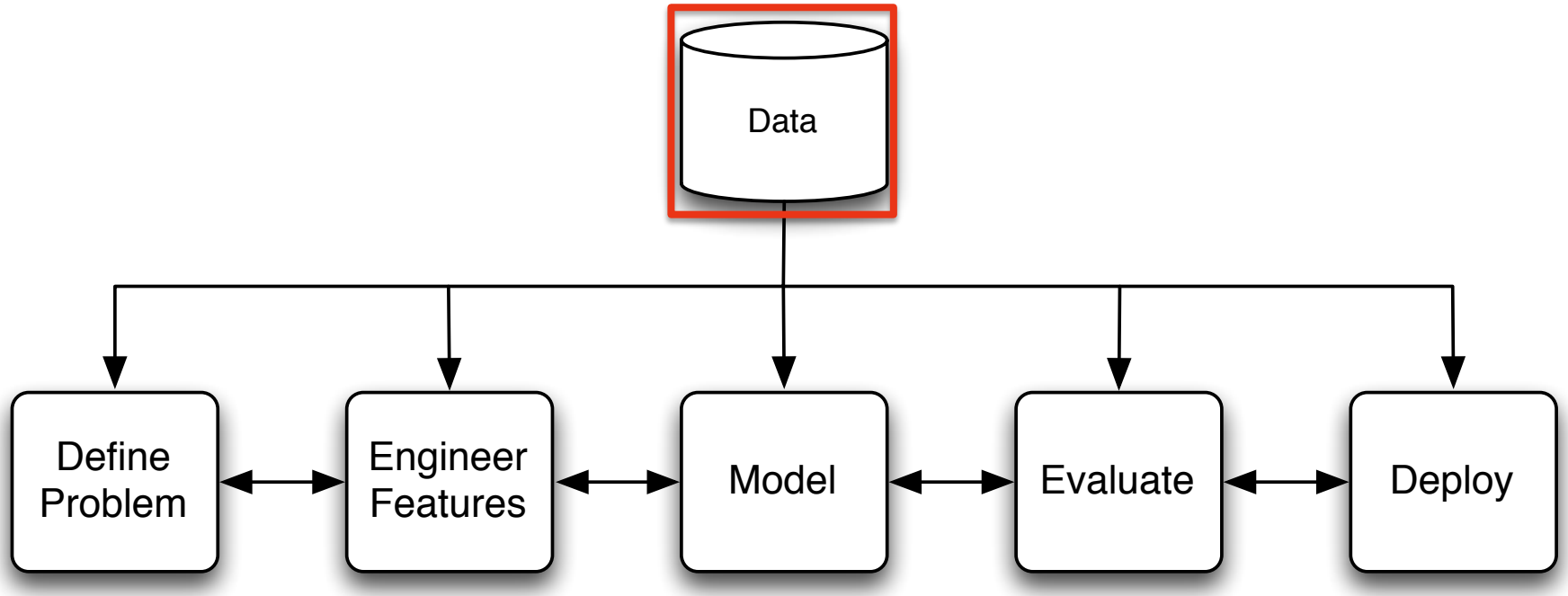
Damballa Labs



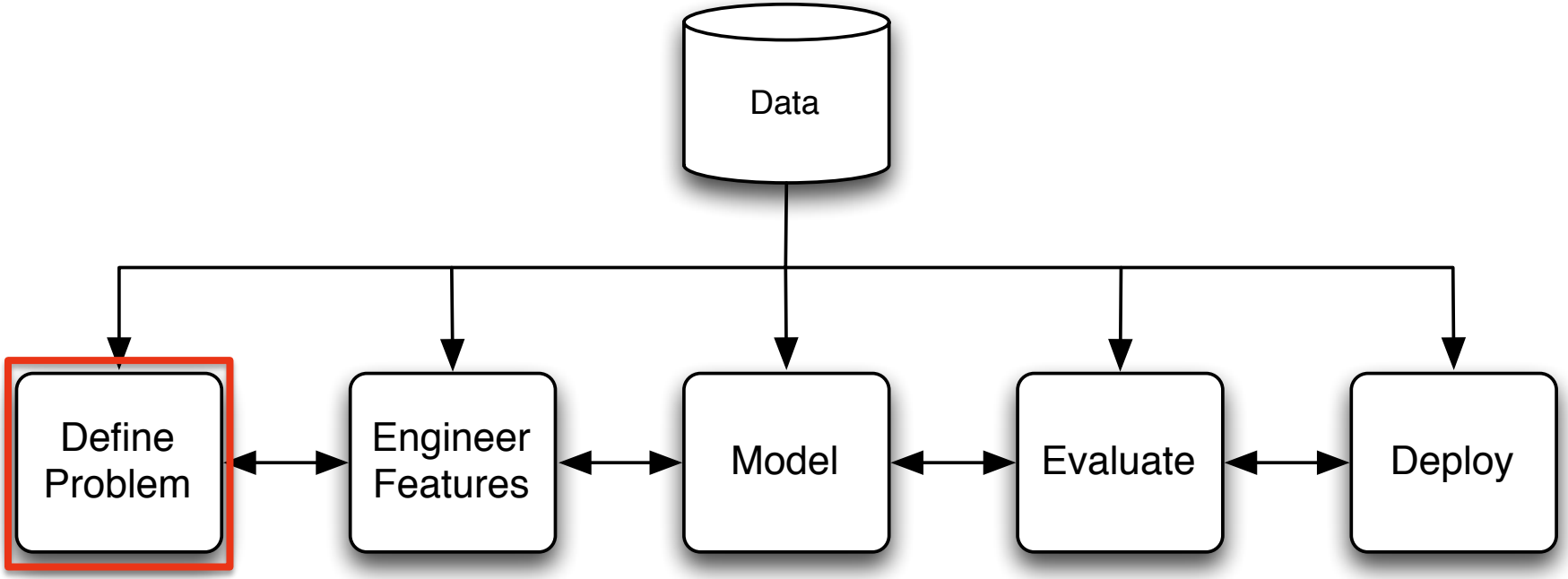
Pratical Machine Learning for Network Security



Pratical Machine Learning for Network Security



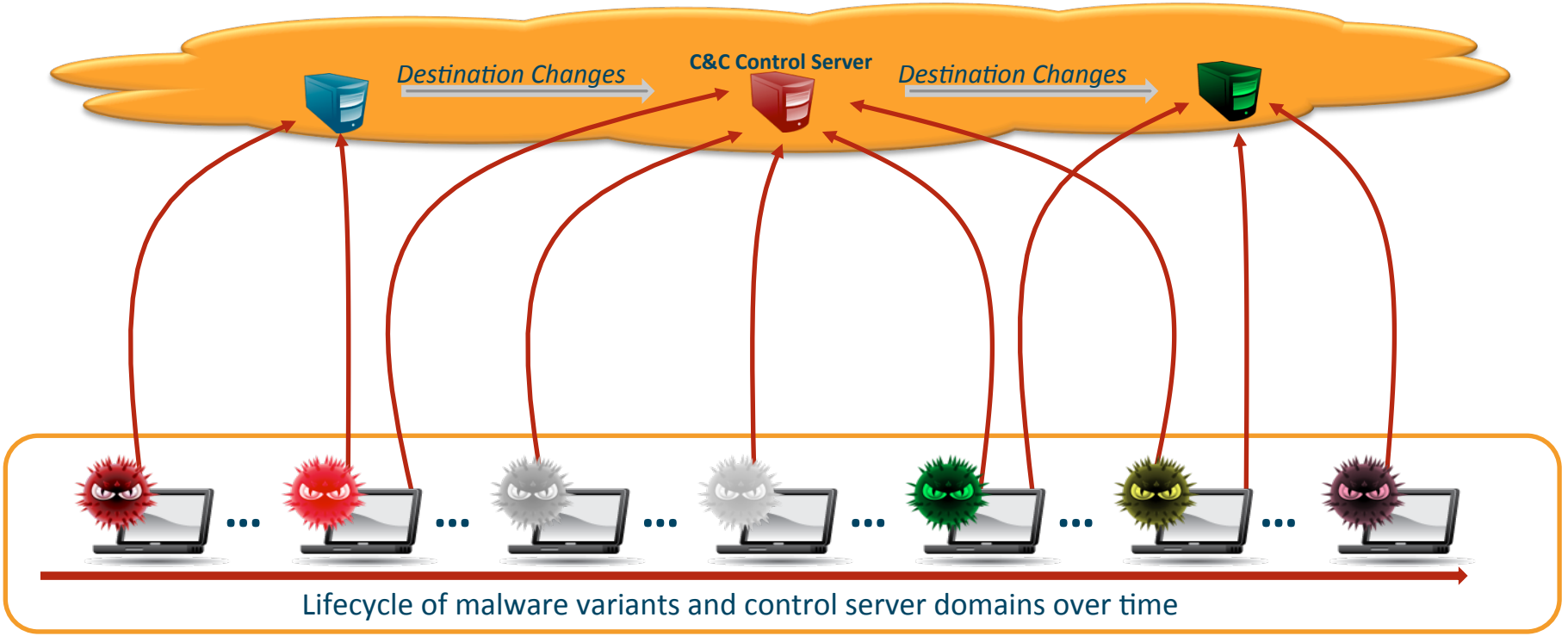
Pratical Machine Learning for Network Security



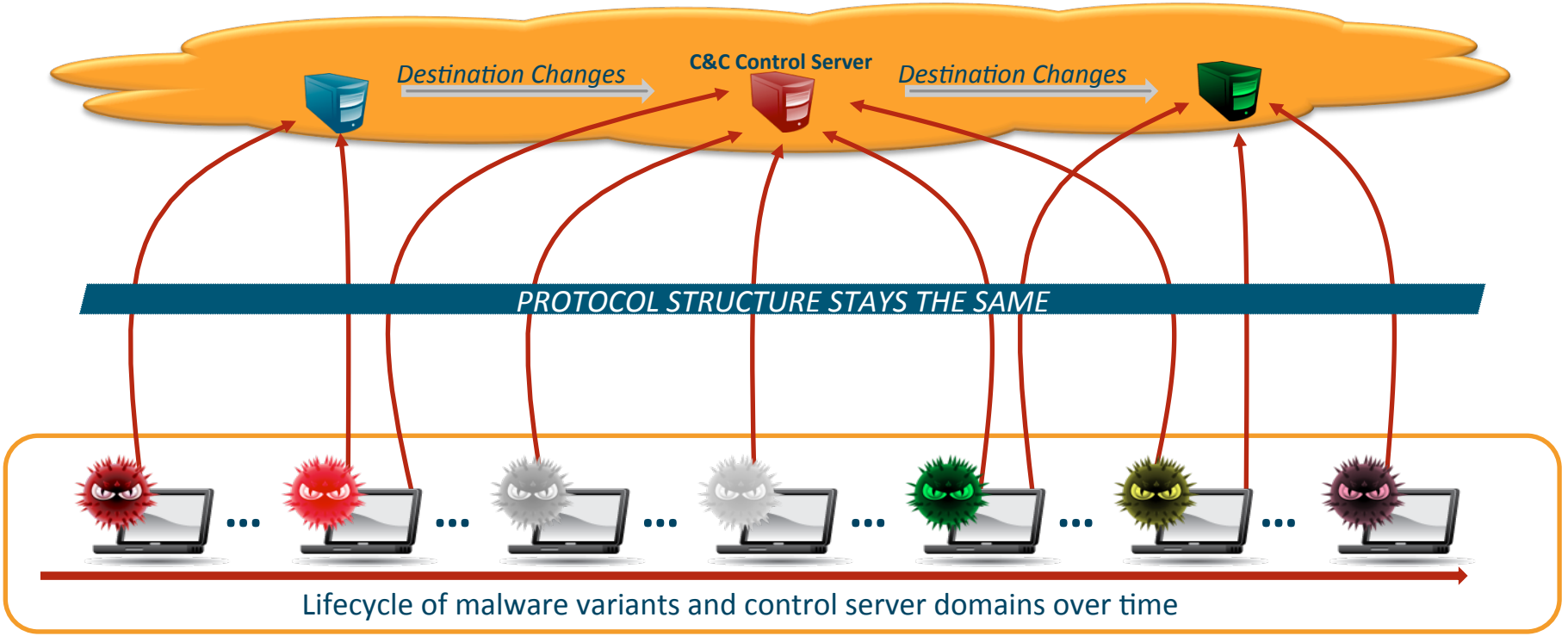
High Level Problem Definition

- Detect hosts infected with malware through observing their network communication.

Malware Command & Control



Malware Command & Control



Defining the Problem – C&C Protocol Detection

- **Task:** recognizing and attributing C&C communication on live networks.
- **Training experience:** packet captures of labeled C&C communication.
- **Performance measurement:** percentage of network communication correctly classified.

DGA-Based Malware

```
qrl89y666z.tang.la
p5ctnvqyd3.myftp.org
5opskttv3y.serveblog.net
tzeh62imx.informatix.com.ru
0zd2bwqqyu.no-ip.info
2ndk2swdma.madhacker.biz
pe4d0t35bs.no-ip.info
5c0x3re4vr.zapto.org
seqkhgd4pj.logout.us
zkycgbn8es.serveblog.net
a4669k3.spacetechnology.net
0few3kd4yv.mooo.info
…
```

Day-0

```
f7865kd.spacetechnology.net
rlq07y626z.tang.la
pf4d9t24bs.no-ip.info
5opskttv3y.serveblog.net
jzek52imx.informatix.com.ru
0zd2bwqqyu.no-ip.info
sgqfhgq2pj.logout.us
9mdk5szdga.madhacker.biz
2c0x3re8vr.zapto.org
qkyrgbn7es.serveblog.net
9fdw3kd4yv.mooo.info
q8ctgvqzd7.myftp.org
…
```

Day-1

```
34ptkssv5y.serveblog.net
a15ctnzqyd3.myftp.org
jh9etzeh5.informatix.com.ru
0zd2bwqqyu.no-ip.info
rkgcgcn8es.serveblog.net
55nk9swffa.madhacker.biz
pb0d3t32bs.no-ip.info
054g3kd4yv.mooo.info
5c0x3re4vr.zapto.org
vb4qkhfd4pj.logout.us
vcrej93.spacetechnology.net
nrl89y666z.tang.la
…
```

Day-2

DGA-Based Malware

qrl89y666z.tang.la
p5ctnvqyd3.myftp.org
5opskttv3y.serveblog.net
tzeh62imx.informatix.com.ru
0zd2bwqqyu.no-ip.info
2ndk2swdma.madhacker.biz
pe4d0t35bs.no-ip.info
5c0x3re4vr.zapto.org
seqkhgd4pj.logout.us
zkycgbn8es.serveblog.net
a4669k3.spacetechnology.net
0few3kd4yv.mooo.info
…

Day-0

f7865kd.spacetechnology.net
rlq07y626z.tang.la
pf4d9t24bs.no-ip.info
5opskttv3y.serveblog.net
jzek52imx.informatix.com.ru
0zd2bwqqyu.no-ip.info
sgqfhgq2pj.logout.us
9mdk5szdga.madhacker.biz
2c0x3re8vr.zapto.org
qkyrgbn7es.serveblog.net
9fdw3kd4yv.mooo.info
q8ctgvqzd7.myftp.org
…

Day-1

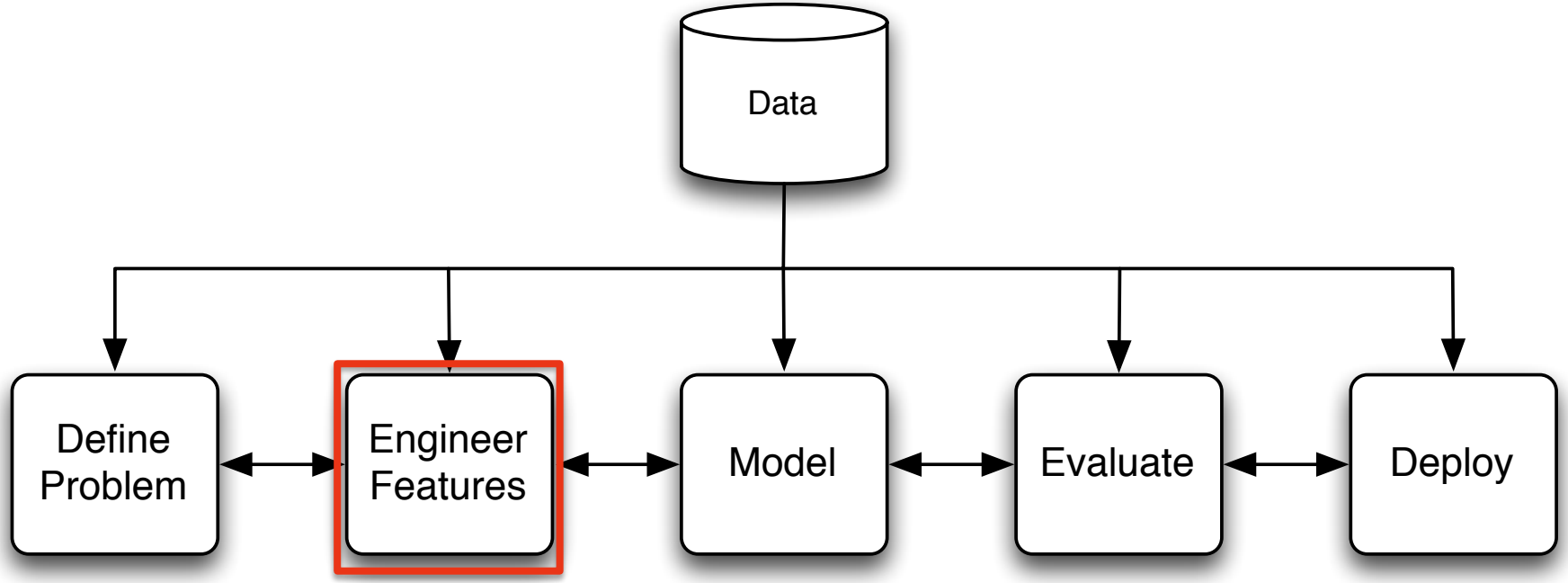
34ptkssv5y.serveblog.net
a15ctnzqyd3.myftp.org
jh9etzeh5.informatix.com.ru
0zd2bwqqyu.no-ip.info
rkgcgcn8es.serveblog.net
55nk9swffa.madhacker.biz
pb0d3t32bs.no-ip.info
054g3kd4yv.mooo.info
5c0x3re4vr.zapto.org
vb4qkhfd4pj.logout.us
vcrej93.spacetechnology.net
nrl89y666z.tang.la
…

Day-2

Defining the Problem – DGA Detection

- **Task:** recognizing and attributing sets of NXDomains to a DGA.
- **Training experience:** labeled sets of NXDomains.
- **Performance measurement:** percentage of NXDomains correctly classified.

Pratical Machine Learning for Network Security



Generalizing C&C Protocol Structure

Request 1:
GET /Ym90bmq=/cnc.php?v=220&cc=IT
Host: www.bot.net
User-Agent: 680e4a9a

Request 2:
GET /bWFsd2F=/cnc.php?v=139&cc=US
Host: www.malwa.re
User-Agent: dae4a661

Generalizing C&C Protocol Structure

Request 1:
GET /Ym90bmq=/cnc.php?v=220&cc=IT
Host: www.bot.net
User-Agent: 680e4a9a

Request 2:
GET /bWFsd2F=/cnc.php?v=139&cc=US
Host: www.malwa.re
User-Agent: dae4a661

Generalized Request 1:
GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>
Host: www.bot.net
User-Agent: <hex,8>

Generalized Request 2:
GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>
Host: www.malwa.re
User-Agent: <hex,8>

Features – Query Names

Generalized Request 1:

GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>

Host: www.bot.net

User-Agent: <hex,8>

Generalized Request 2:

GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>

Host: www.malwa.re

User-Agent: <hex,8>

Features – Query Data Types & Lengths

Generalized Request 1:

GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>

Host: www.bot.net

User-Agent: <hex,8>

Generalized Request 2:

GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>

Host: www.malwa.re

User-Agent: <hex,8>

Features – Path

Generalized Request 1:

GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>

Host: www.bot.net

User-Agent: <hex,8>

Generalized Request 2:

GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>

Host: www.malwa.re

User-Agent: <hex,8>

Features – Headers

Generalized Request 1:

GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>

Host: www.bot.net

User-Agent: <hex,8>

Generalized Request 2:

GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>

Host: www.malwa.re

User-Agent: <hex,8>

Features – IP addresses hosting domain

Generalized Request 1:

GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>

Host: www.bot.net

User-Agent: <hex,8>

Generalized Request 2:

GET /<base64,8>/cnc.php?v=<int,3>&cc=<str,2>

Host: www.malwa.re

User-Agent: <hex,8>

DGA Feature Engineering

DGA 1

```
fgalu-xixes.com
fgala-kekif.com
fgaky-xafog.com
fgaku-kynam.com
fgaji-megaz.com
fvopyv-uzir.ru
fvopym-ivef.ru
fvopoz-ekir.ru
fvoniw-aker.ru
fvonek-uwif.ru
fvomyz-ymaz.ru
fvomyd-otir.ru
…
```

DGA 2

```
qrl89y666z.tang.la
p5ctnvqyd3.myftp.org
5opskttv3y.serveblog.net
tzeh62ismx.informatix.com.ru
0zd2bwqqyu.no-ip.info
2ndk2swdma.madhacker.biz
pe4d0t35bs.no-ip.info
5c0x3re4vr.zapto.org
seqkhgd4pj.logout.us
zkycgbn8es.serveblog.net
a41u669vk3.spacetechnology.net
0few3kd4yv.mooo.info
…
```

Features - *n*-gram

bigram

fgalu-xixes.com
fgalu-xixes.com
fgalu-xixes.com
fgalu-xixes.com
fgalu-xixes.com
fgalu-xixes.com
fgalu-xixes.com
fgalu-xixes.com
fgalu-xixes.com
fgalu-xixes.com

DGA 1

fgalu-xixes.com
fgala-kekif.com
fgaky-xafog.com
fgaku-kynam.com
fgaji-megaz.com
fvopyv-uzir.ru
fvopym-ivef.ru
fvopoz-ekir.ru
fvoniw-aker.ru
fvonek-uwif.ru
fvomyz-ymaz.ru
fvomyd-otir.ru
…

DGA 2

qrl89y666z.tang.la
p5ctnvqyd3.myftp.org
5opskttv3y.serveblog.net
tzeh62ismx.informatix.com.ru
0zd2bwqqyu.no-ip.info
2ndk2swdma.madhacker.biz
pe4d0t35bs.no-ip.info
5c0x3re4vr.zapto.org
seqkhgd4pj.logout.us
zkycgbn8es.serveblog.net
a41u669vk3.spacetechnology.net
0few3kd4yv.mooo.info
…

trigram

qrl89y666z.tang.la
qrl89y666z.tang.la
qrl89y666z.tang.la
qrl89y666z.tang.la
qrl89y666z.tang.la
qrl89y666z.tang.la
qrl89y666z.tang.la
qrl89y666z.tang.la

Features – Entropy

Lower

DGA 1	DGA 2
fgalu-xixes.com	qrl89y666z.tang.la
fgala-kekif.com	p5ctnvqyd3.myftp.org
fgaky-xafog.com	5opskttv3y.serveblog.net
fgaku-kynam.com	tzeh62ismx.informatix.com.ru
fgaji-megaz.com	0zd2bwqqyu.no-ip.info
fvopyv-uzir.ru	2ndk2swdma.madhacker.biz
fvopym-ivef.ru	pe4d0t35bs.no-ip.info
fvopoz-ekir.ru	5c0x3re4vr.zapto.org
fvoniw-aker.ru	seqkhgd4pj.logout.us
fvonek-uwif.ru	zkycgbn8es.serveblog.net
fvomyz-ymaz.ru	a41u669vk3.spacetechnology.net
fvomyd-otir.ru	0few3kd4yv.mooo.info
…	…

Higher

Features – Structural

DGA 1

fgalu-xixes.com
fgala-kekif.com
fgaky-xafog.com
fgaku-kynam.com
fgaji-megaz.com
fvopyv-uzir.ru
fvopym-ivef.ru
fvopoz-ekir.ru
fvoniw-aker.ru
fvonek-uwif.ru
fvomyz-ymaz.ru
fvomyd-otir.ru
…

DGA 2

qrl89y666z.tang.la
p5ctnvqyd3.myftp.org
5opskttv3y.serveblog.net
tzeh62ismx.informatix.com.ru
0zd2bwqqyu.no-ip.info
2ndk2swdma.madhacker.biz
pe4d0t35bs.no-ip.info
5c0x3re4vr.zapto.org
seqkhgd4pj.logout.us
zkycgbn8es.serveblog.net
a41u669vk3.spacetechnology.net
0few3kd4yv.mooo.info
…

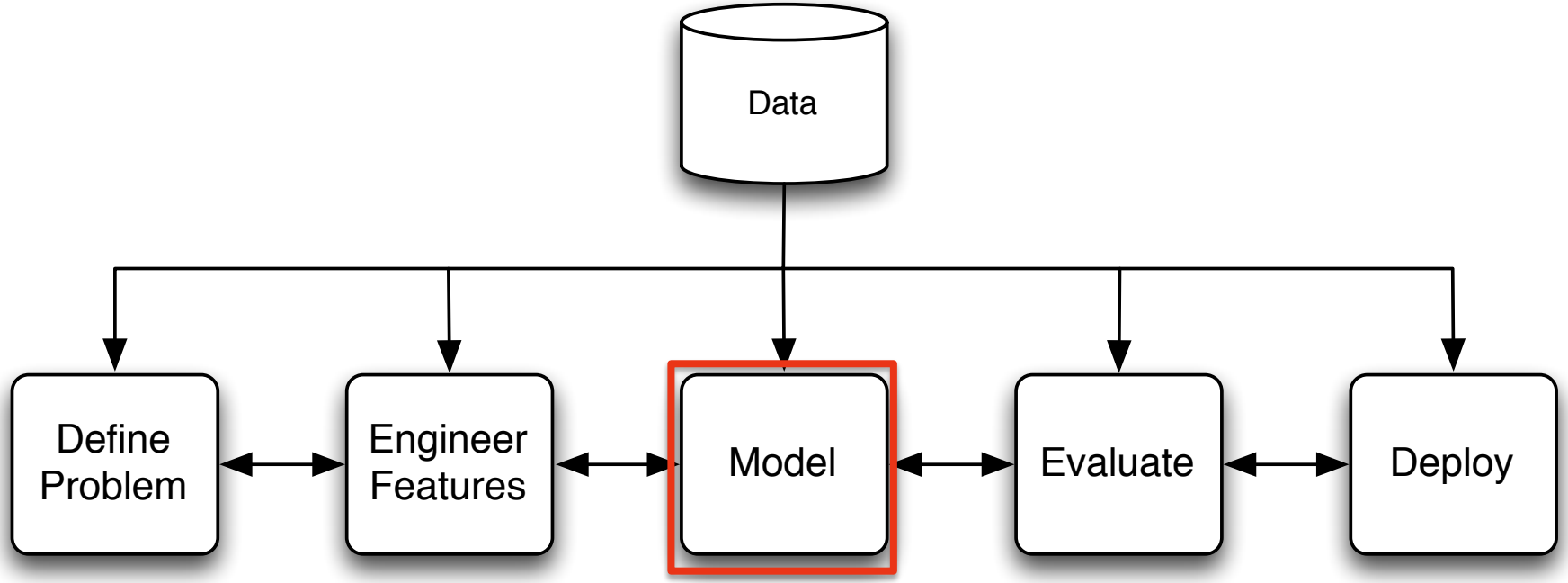
Features – NXDomain/Client

	H_1	H_2	H_3	…	H_m
NX_1	1	**1**	0	0	**1**
NX_2	0	**1**	0	0	**1**
…	0	**0**	1	0	**0**
NX_n	0	**1**	0	1	**1**

Determining important features

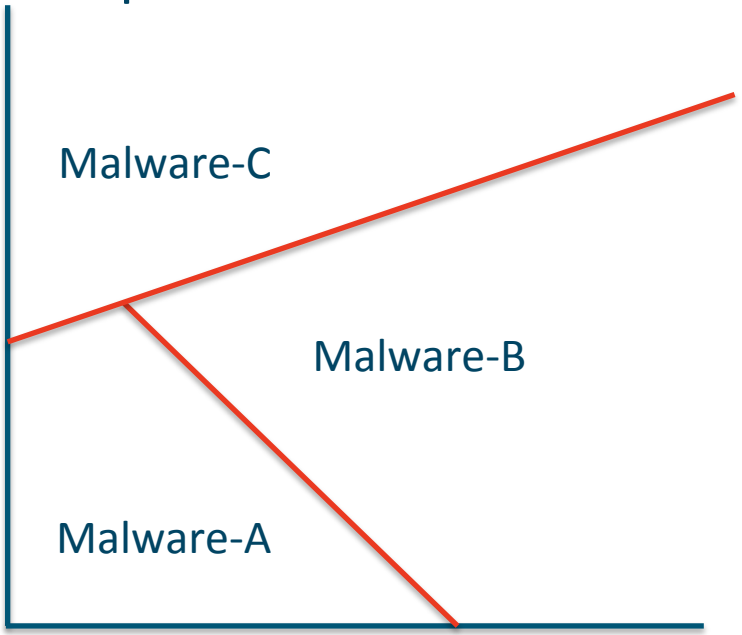
- Try all combination of features (2^n).
- Forward selection.
- Backwards selection.

Pratical Machine Learning for Network Security

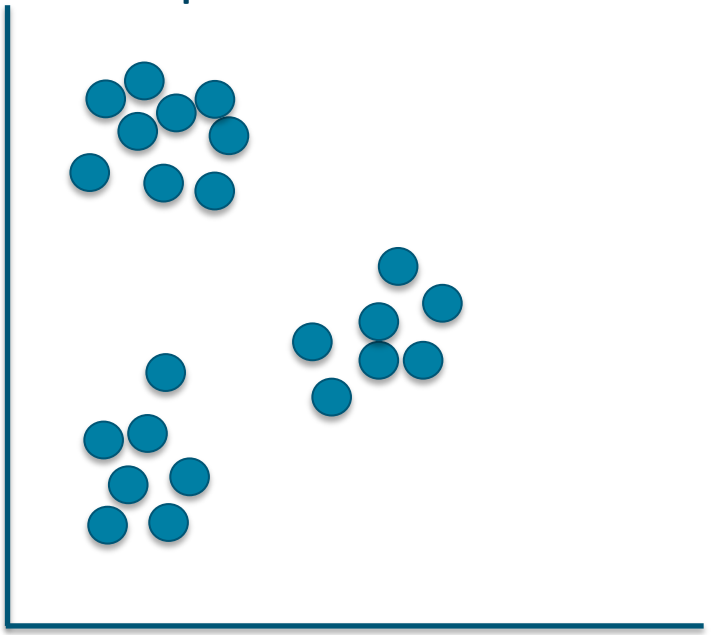


Learning: Supervised vs. Unsupervised

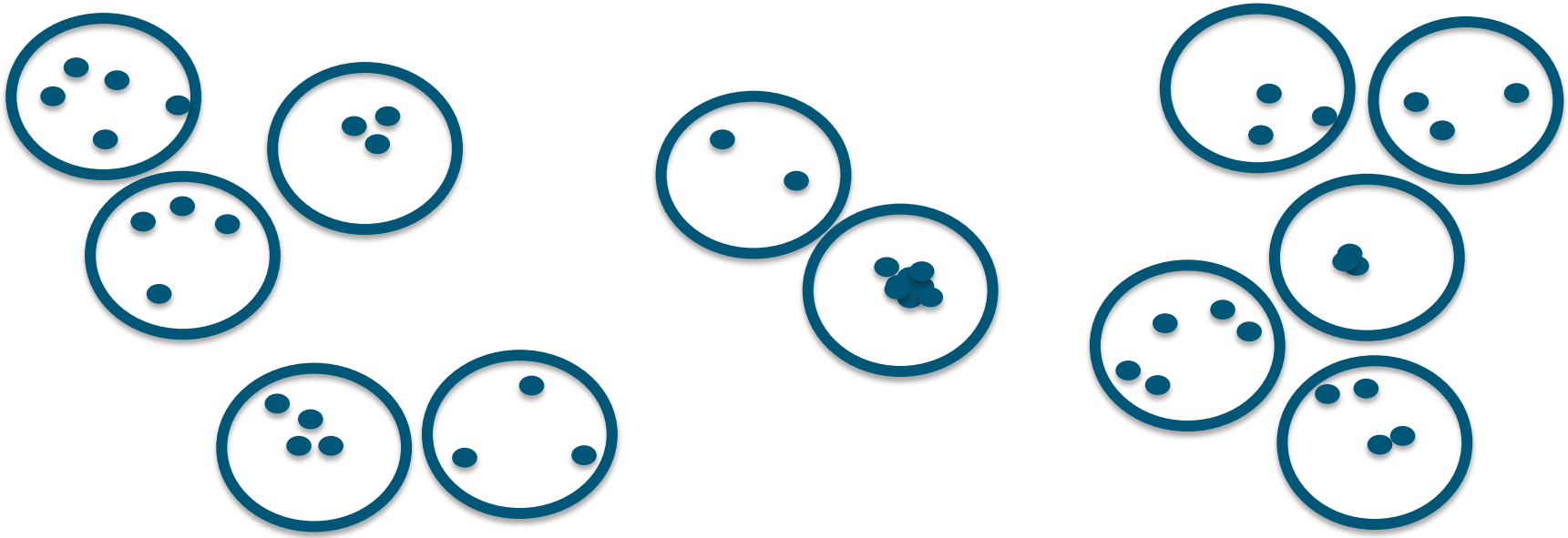
Supervised



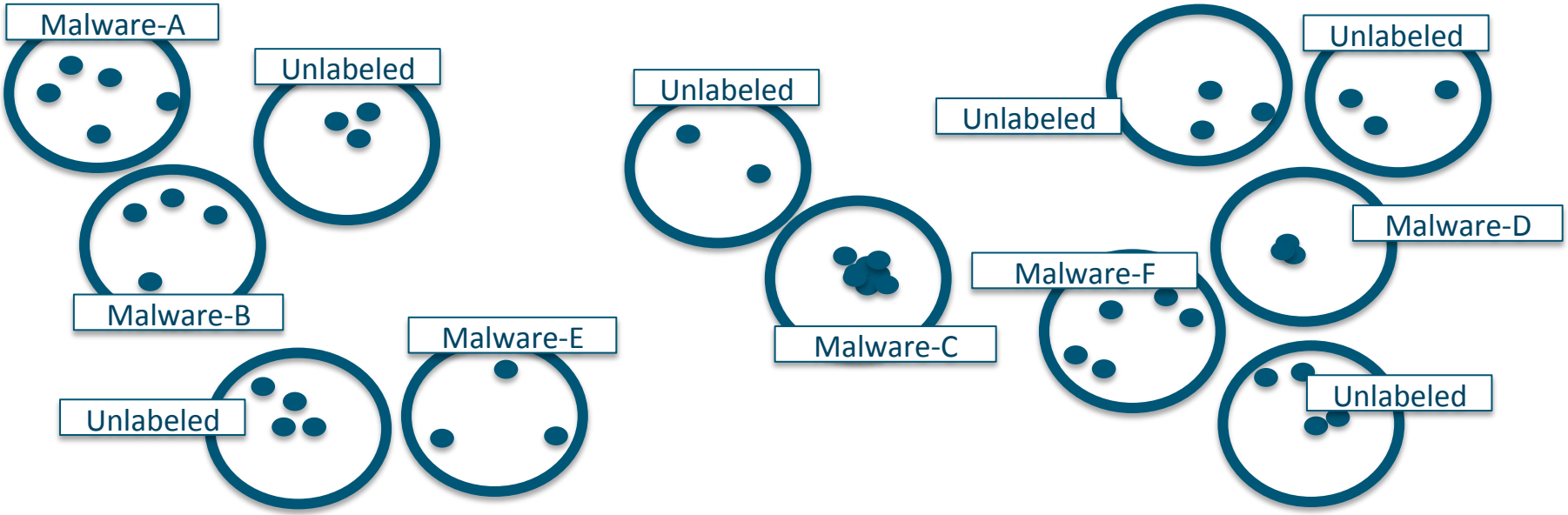
Unsupervised



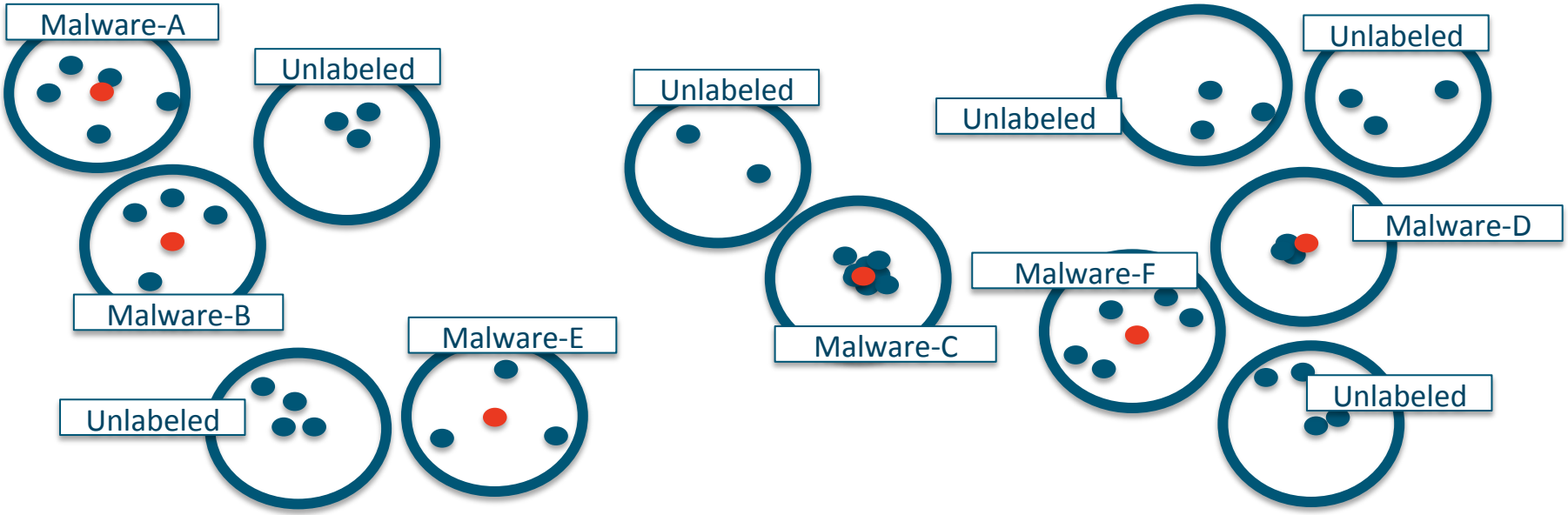
Unsupervised Learning - Clustering HTTP Requests



Unsupervised Learning - Clustering HTTP Requests



Unsupervised Learning - Clustering HTTP Requests



C&C Protocol Detection

- Similarity
  - Measures likeness
  - CPT specific
- Specificity
  - Measures uniqueness
  - Network specific

Input: req, CPT

Similarity: $s(\text{req}_i, \text{CPT}_i)$, for each component i

Specificity: $\sigma(\text{req}_i, \text{CPT}_i)$, for each component i

Match-Score: $f(\text{sim}, \text{spec})$

If Match-Score > Θ:
    return C&C Request

Unsupervised Learning - Clustering DGAs

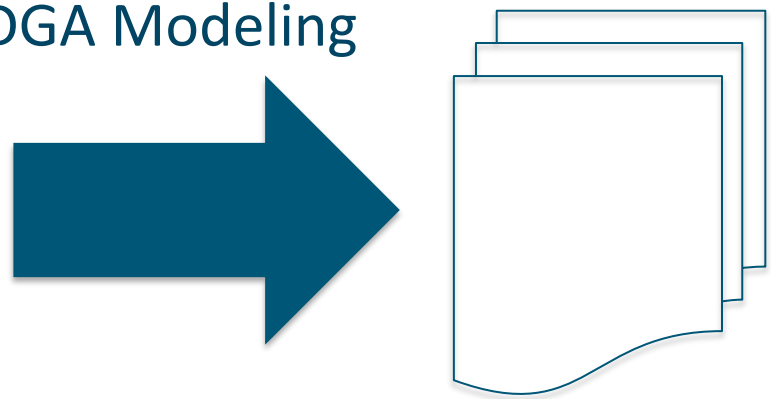
DGA 1	DGA 2
fgalu-xixes.com	qrl89y666z.tang.la
fgala-kekif.com	p5ctnvqyd3.myftp.org
fgaky-xafog.com	5opskttv3y.serveblog.net
fgaku-kynam.com	tzeh62ismx.informatix.com.ru
fgaji-megaz.com	0zd2bwqqyu.no-ip.info
fvopyv-uzir.ru	2ndk2swdma.madhacker.biz
fvopym-ivef.ru	pe4d0t35bs.no-ip.info
fvopoz-ekir.ru	5c0x3re4vr.zapto.org
fvoniw-aker.ru	seqkhgd4pj.logout.us
fvonek-uwif.ru	zkycgbn8es.serveblog.net
fvomyz-ymaz.ru	a41u669vk3.spacetechnology.net
fvomyd-otir.ru	0few3kd4yv.mooo.info
…	…

	H_1	H_2	H_3	…	H_m
NX_1	1	**1**	0	0	**1**
NX_2	0	**1**	0	0	**1**
…	0	**0**	1	0	**0**
NX_n	0	**1**	0	1	**1**

Supervised Learning – Modeling DGAs

Malware-A	Malware-B
fgalu-xixes.com	qrl89y666z.tang.la
fgala-kekif.com	p5ctnvqyd3.myftp.org
fgaky-xafog.com	5opskttv3y.serveblog.net
fgaku-kynam.com	tzeh62ismx.informatix.com.ru
fgaji-megaz.com	0zd2bwqqyu.no-ip.info
fvopyv-uzir.ru	2ndk2swdma.madhacker.biz
fvopym-ivef.ru	pe4d0t35bs.no-ip.info
fvopoz-ekir.ru	5c0x3re4vr.zapto.org
fvoniw-aker.ru	seqkhgd4pj.logout.us
fvonek-uwif.ru	zkycgbn8es.serveblog.net
fvomyz-ymaz.ru	a41u669vk3.spacetechnology.net
fvomyd-otir.ru	0few3kd4yv.mooo.info
…	…

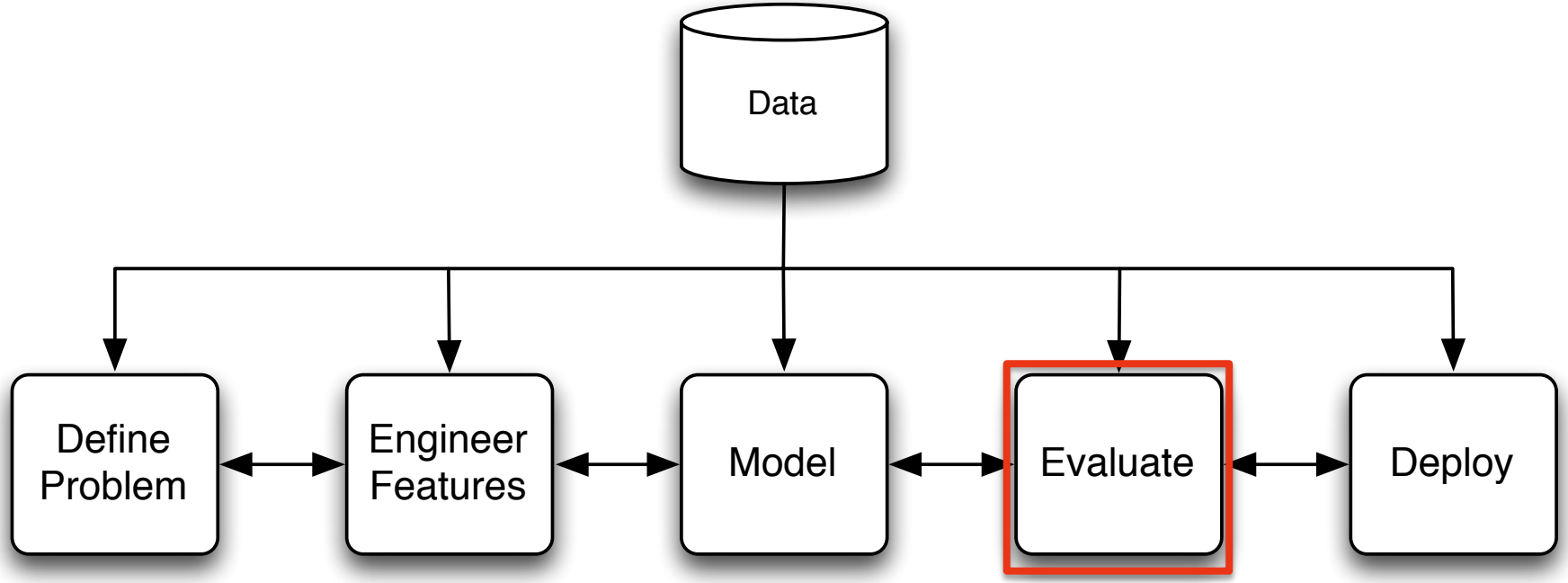
DGA Modeling



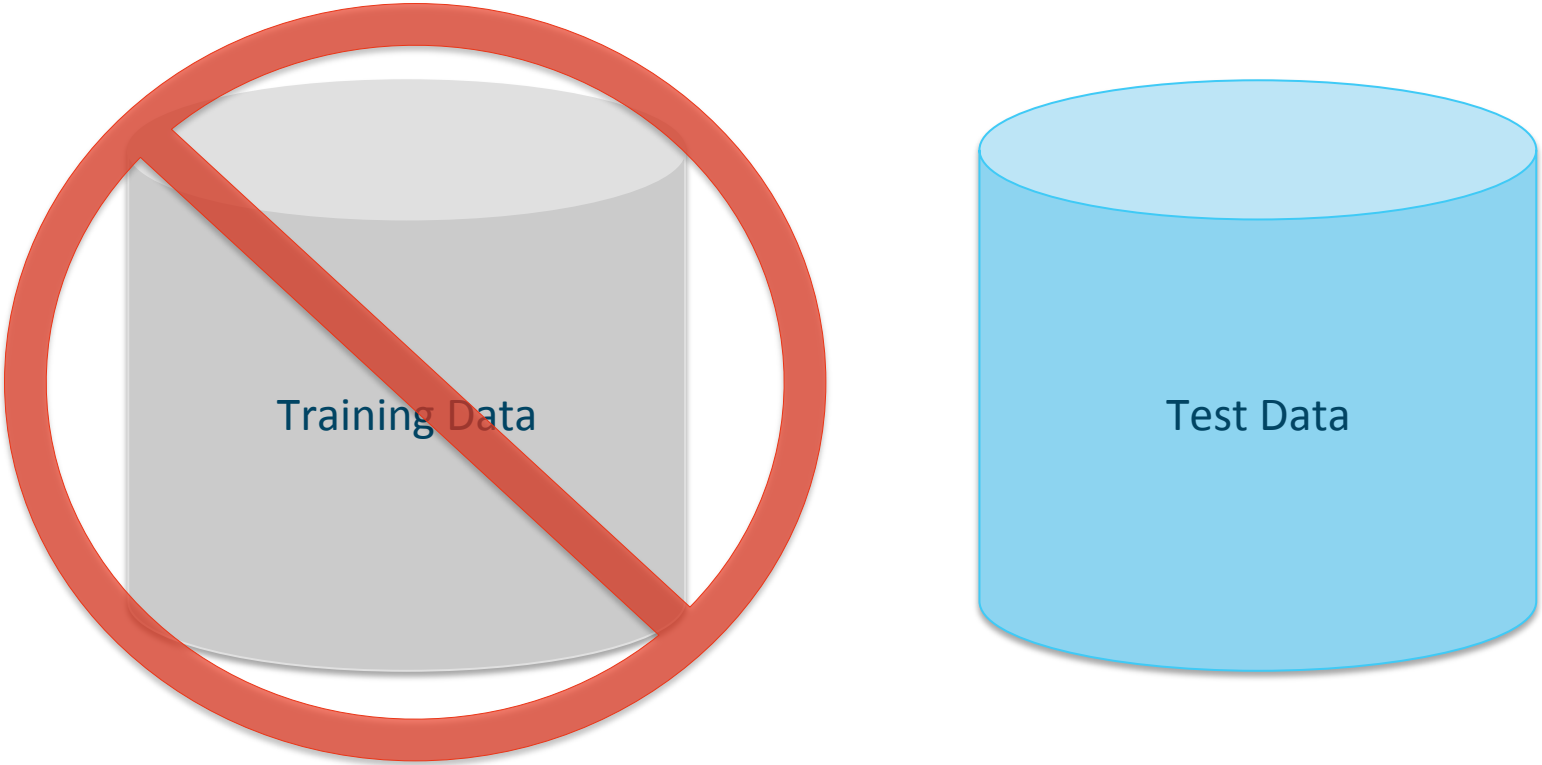
Modeling Tools

- **Scikit-learn**
  - Collection of machine learning algorithms (Python).
  - http://scikit-learn.org/stable/
- **Weka**
  - Collection of machine learning algorithms (Java).
  - http://www.cs.waikato.ac.nz/ml/weka/
- **R**
  - Language and environment for statistical computing and graphics.
  - http://www.r-project.org/

Pratical Machine Learning for Network Security



Evaluation Data

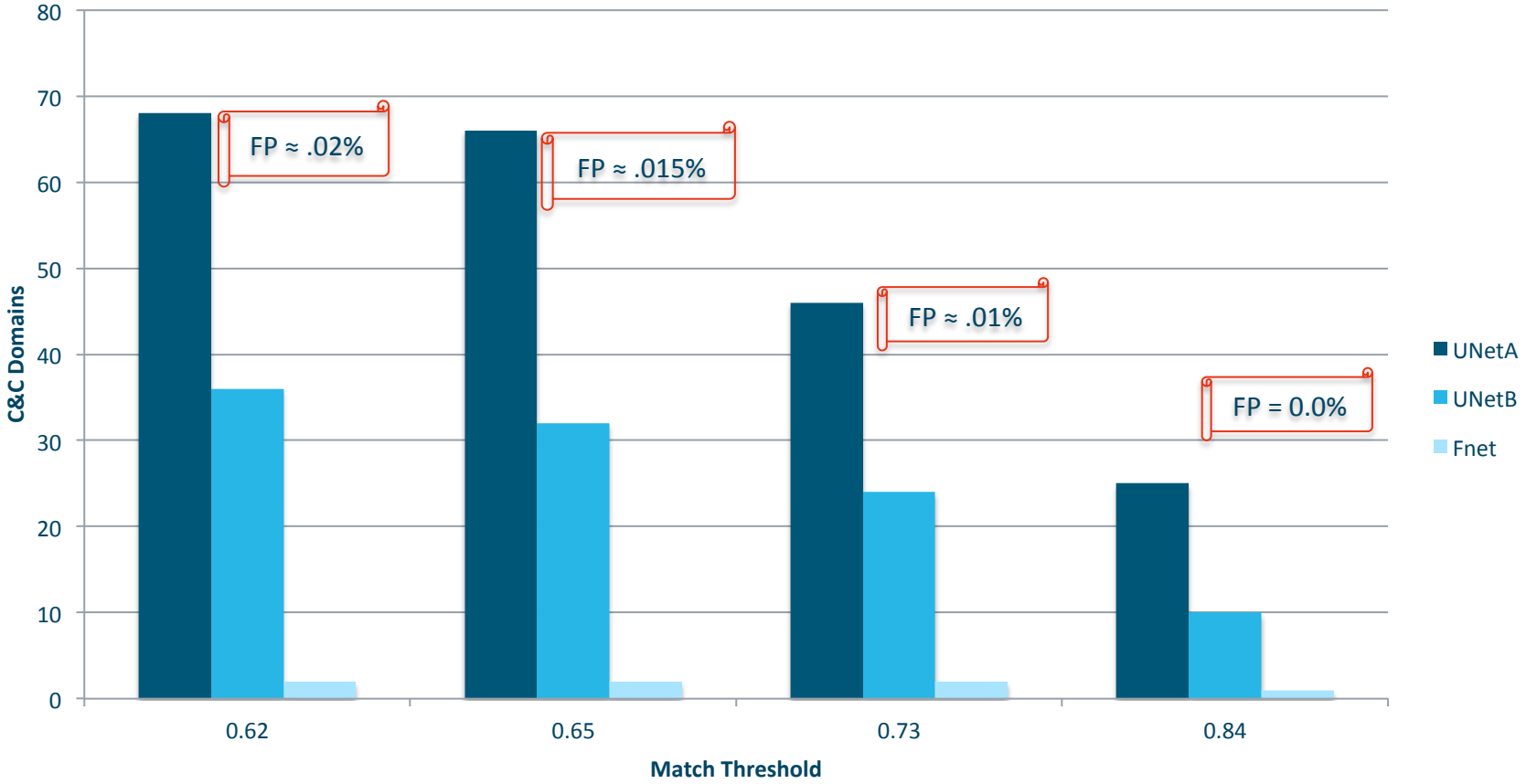


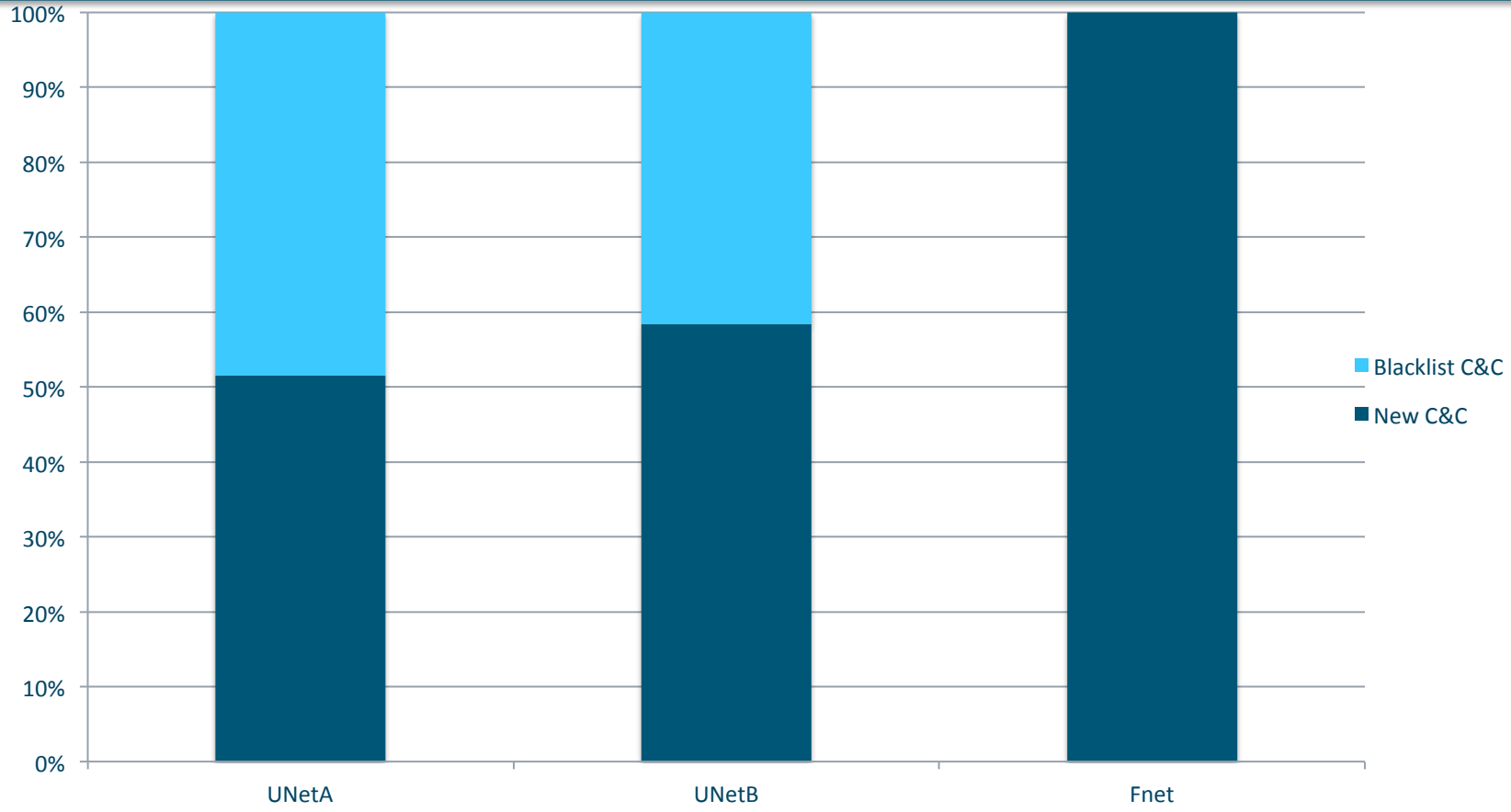
C&C Evaluation Deployment Networks

	UNetA	UNetB	FNet
Distinct Src IPs	7,893	27,340	7,091
HTTP Requests	34,871,003	66,298,395	58,019,718
Distinct Domains	149,481	238,014	113,778

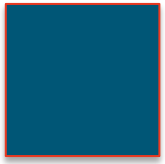
- Evaluation ran for two weeks.
- CPTs updated daily beginning two weeks prior to evaluation.

Network Deployment Results

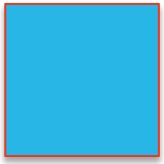



100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
UNetA
UNetB
Fnet
Blacklist C&C
New C&C

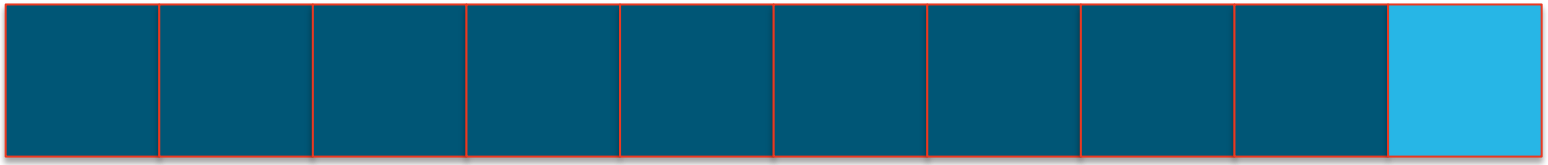
k-fold Cross Validation



Training

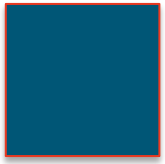


Testing

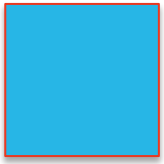


$k = 10$

k-fold Cross Validation



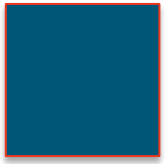
Training



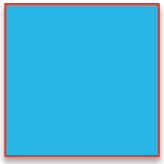
Testing

$k = 10$

k-fold Cross Validation



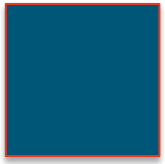
Training



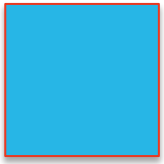
Testing

$k = 10$

k-fold Cross Validation



Training



Testing

$k = 10$

DGA Classifier - 10-fold Cross Validation

Botnet	TP Rate	FP Rate
Bobax	99%	0%
Conficker	99%	0.1%
Sinowal	100%	0%
Murofet	99%	0.2%
Benign	99%	0.1%

DGA Clustering – ISP Deployment

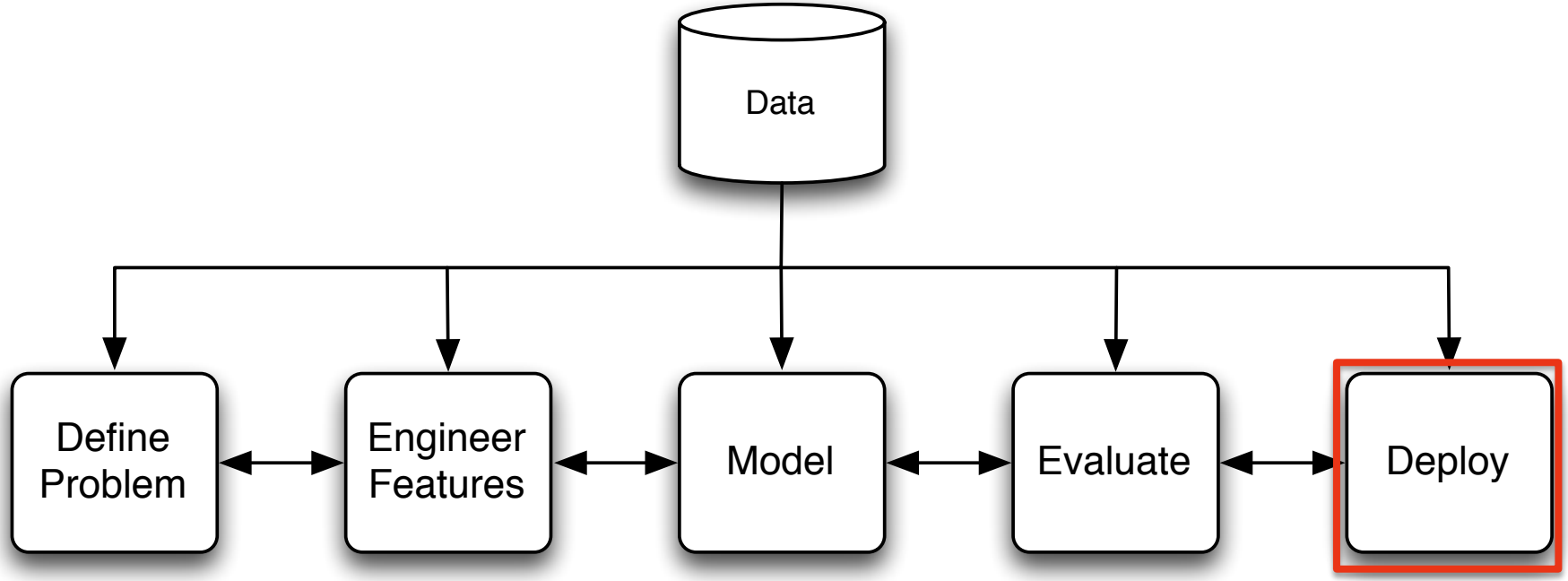
- Six confirmed DGA-based malware
- Six new DGAs for which no malware family (at discovery)

Malware Family	First Seen	Population on Discovery
Shiz/Simda-C [32]	03/20/11	37
Bamital [11]	04/01/11	175
BankPatch [5]	04/01/11	28
Expiro.Z [8]	04/30/11	7
Bonnana [41]	08/03/11	24
Zeus.v3 [25]	09/15/11	39
New-DGA-v1	01/11/10	12
New-DGA-v2	01/18/11	10
New-DGA-v3	02/01/11	18
New-DGA-v4	03/05/11	22
New-DGA-v5	04/21/11	5
New-DGA-v6	11/20/11	10

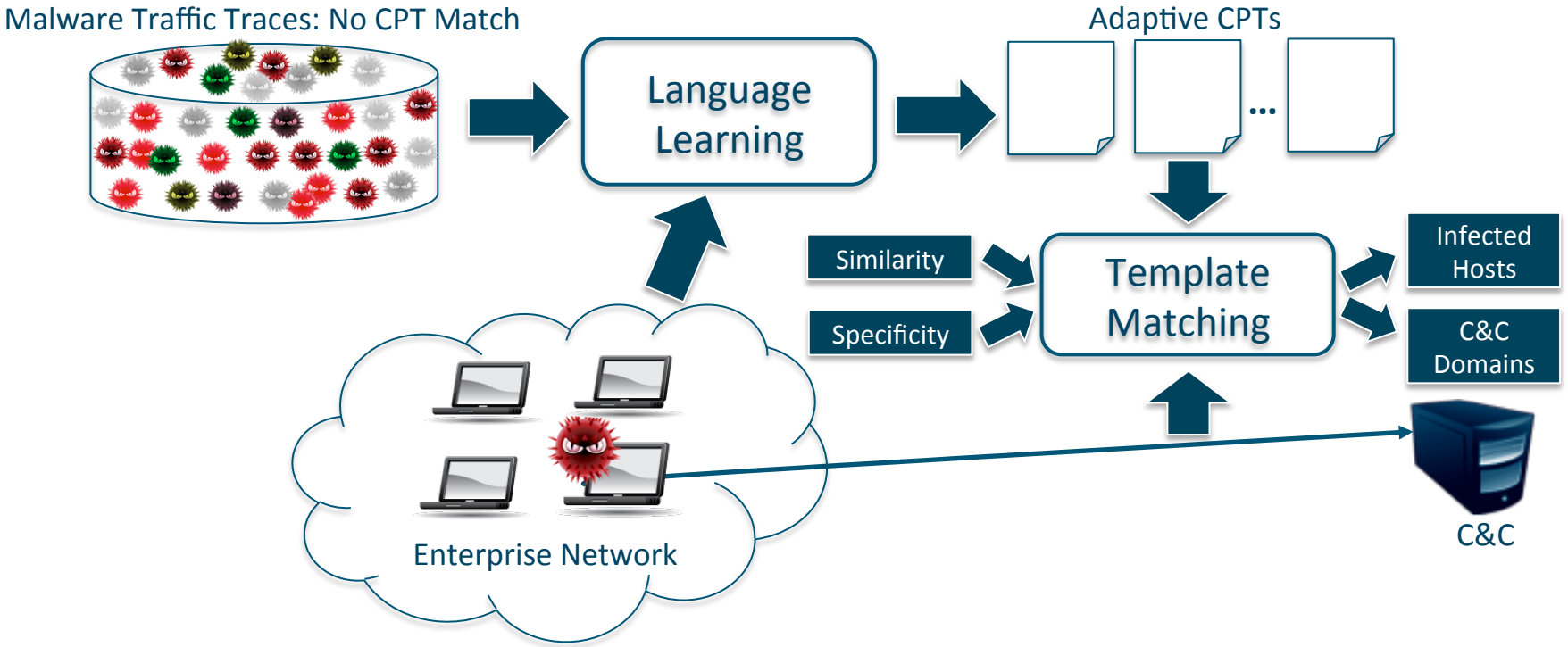
New-DGA-v1	New-DGA-v2	New-DGA-v3
71f9d3d1.net	clfnoooqfpdc.com	uwhornfrqsdbrbnbuhjt.com
a8459681.com	slsleujrrzwx.com	epmsgxuotsciklvywmck.com
a8459681.info	qzycprhfiwfb.com	nxmglieidfsdolcakggk.com
a8459681.net	uvphgewngjiq.com	ieheckbkkkoibskrqana.com
1738a9aa.com	gxnbtlvvwmyg.com	qabgwxmkqdeixsqavxhr.com
1738a9aa.info	wdlmurglkuxb.com	gmjvfbhfcfkfyotdvbtv.com
1738a9aa.net	zzopaahxctfh.com	sajltlsbigtfexpxvsri.com
84c7e2a3.com	bzqbcftfcrqf.com	uxyjfflvoqoephfywjcq.com
84c7e2a3.info	rjvmrkkycfuh.com	kantifyosseefhdgilha.com
84c7e2a3.net	itzbkyunmzfv.com	lmklwkkrficnnqugqlpj.com

New-DGA-v4	New-DGA-v5	New-DGA-v6
semk1cquvjufayg02orednzdfg.com	zpdyaislnu.net	lymylorozig.eu
invfgg4szr22sbjbmdqm51pdtf.com	vvbmjfxpyi.net	lyvejujolec.eu
0vqbqcuqdv0i1fadodtm5iumye.com	oisbyccilt.net	xuxusujenes.eu
np1r0vnqjr3vbs3c3iqyuwe3vf.com	vgkblzdsde.net	gacezobeqon.eu
s3fhkbdu4dmc001tmxskleeqrf.com	bxrvftzvoc.net	tufecagemyl.eu
gup1iapsm2xiedyefet21sxete.com	dlftozdnxn.net	lyvitexemod.eu
y5rk0hgujfgo0t4sfers2xolte.com	gybszkmpse.net	mavulymupiv.eu
me5oclqrfano4z0mx4qsbpdufc.com	dycsmcfwwa.net	jenokirifux.eu
jwhnr2uu3zp0ep40cttq3oyeed.com	dpwxwmkbxl.net	fotyriwavix.eu
ja4baqnv02qoxlsjxqrszdziwb.com	ttbkuogzum.net	vojugycavov.eu

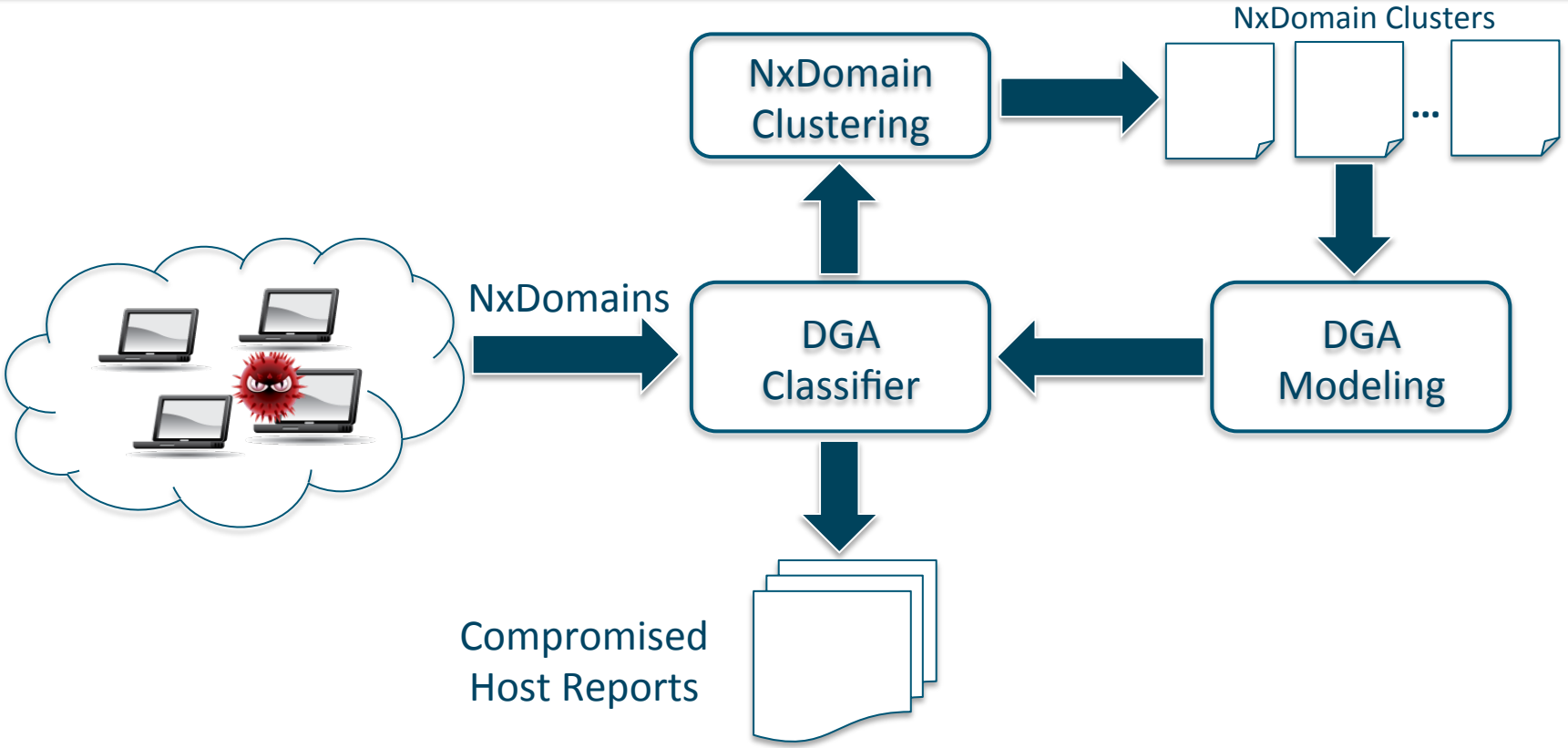
Pratical Machine Learning for Network Security



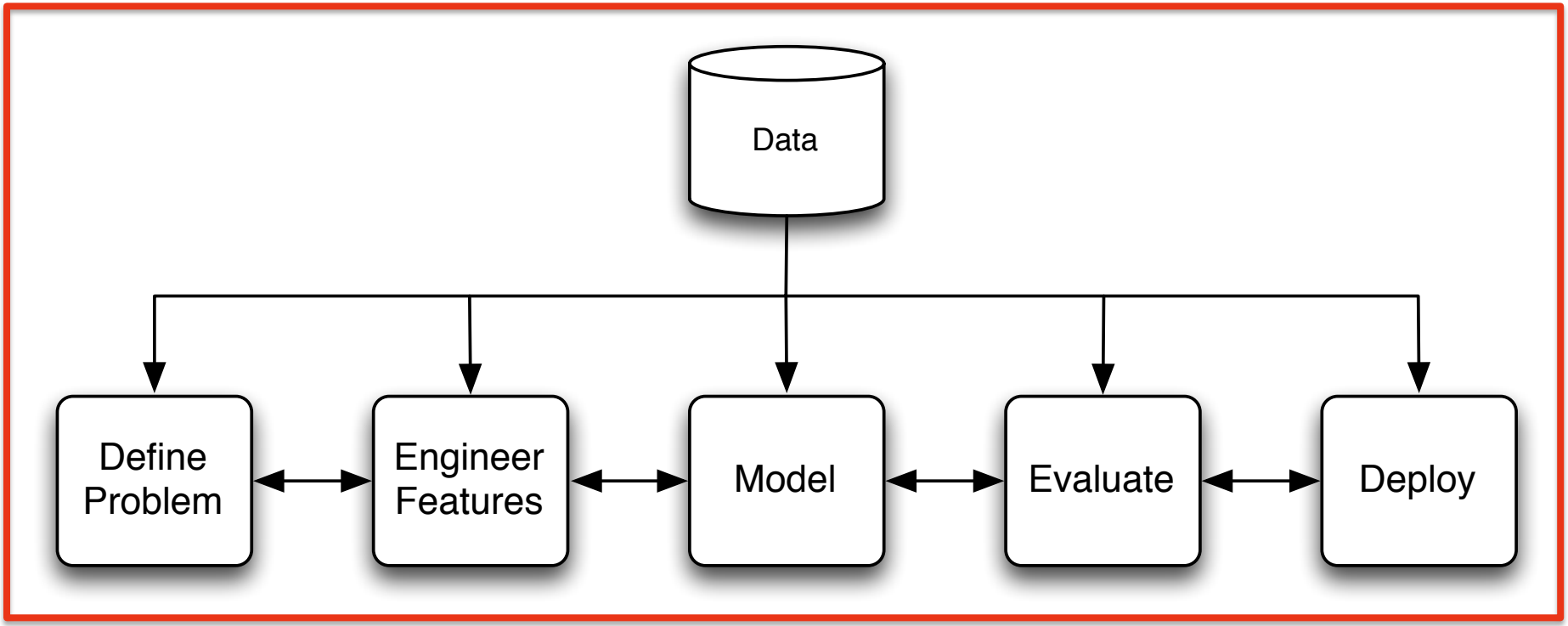
C&C Protocol Detection Deployment



DGA Detection Deployment



Pratical Machine Learning for Network Security



Thank You!

C&C Protocol Detection

Terry Nelms, Roberto Perdisci, and Mustaque Ahamad. 2013. ExecScent: mining for new C&C domains in live networks with adaptive control protocol templates. In Proceedings of the 22nd USENIX conference on Security (SEC'13). USENIX Association, Berkeley, CA, USA, 589-604.

DGA Detection

Manos Antonakakis, Roberto Perdisci, Yacin Nadji, Nikolaos Vasiloglou, Saeed Abu-Nimeh, Wenke Lee, and David Dagon. 2012. From throw-away traffic to bots: detecting the rise of DGA-based malware. In Proceedings of the 21st USENIX conference on Security symposium (Security'12). USENIX Association, Berkeley, CA, USA, 24-24.